

## Chapter 5

# Reward

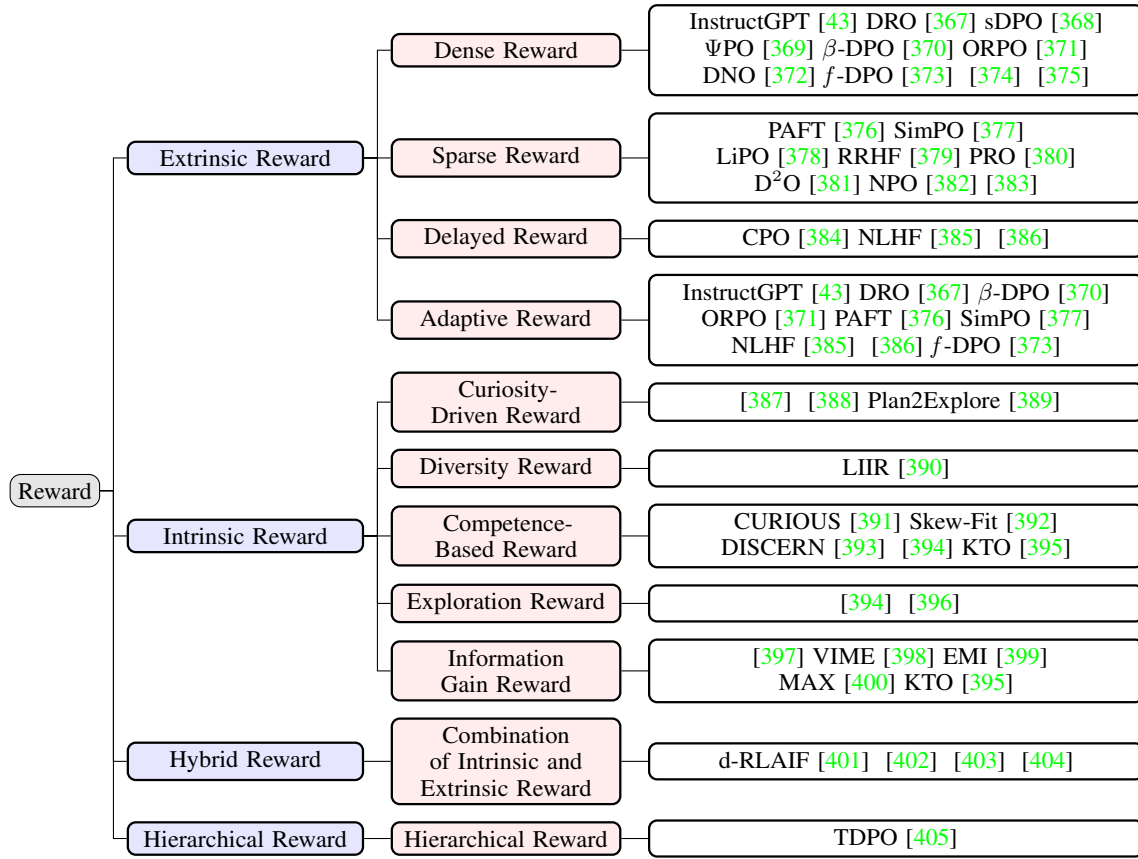


Figure 5.1: Illustrative Taxonomy of Reward system

Rewards help the agent distinguish between beneficial and detrimental actions, shaping its learning process and influencing its decision-making. This chapter first introduces common reward substances in the human body and the corresponding reward pathways. Then, the reward paradigm under the agent and the different methods involved are defined. In the discussion section, the influence relationship between other modules is described, and the existing methods are summarized, then the problems that need to be solved in the future and the optimization directions are discussed.

Table 5.1: The comparison of human common reward pathways.

Reward Pathway	Neurotransmitter	Mechanism
<b>Mesolimbic pathway</b> [406]	Dopamine	Dopaminergic neurons in the ventral tegmental area (VTA) extend projections to the nucleus accumbens, where they release dopamine to regulate reward-related signaling. Dopamine diffuses across the synaptic cleft and binds to dopamine receptors—primarily D1-like (excitatory via Gs proteins, increasing cAMP) and D2-like (inhibitory via Gi proteins, reducing cAMP)—thereby modulating reward, motivation, and reinforcement.
<b>Mesocortical pathway</b> [407]	Dopamine	Dopaminergic projections from the VTA reach the prefrontal cortex (PFC). Here, dopamine binds to its receptors to influence cognitive functions such as decision-making, working memory, and emotional regulation, all of which contribute to evaluating and anticipating rewards.
<b>Nigrostriatal pathway</b> [407]	Dopamine	Dopamine’s action on D1 and D2 receptors in the striatum helps shape both motor routines and reward-related behaviors.
<b>Locus coeruleus</b> [408]	Norepinephrine	Neurons in the locus coeruleus release norepinephrine to widely distributed targets across the brain. At synapses, norepinephrine binds to adrenergic receptors ( $\alpha$ and $\beta$ subtypes), modulating neuronal excitability, arousal, attention, and stress responses. These modulatory effects can indirectly influence reward processing and decision-making circuits.
<b>Glutamatergic projection</b> [409]	Glutamate	Upon releasing into the synaptic cleft, glutamate binds to both ionotropic receptors (such as AMPA and NMDA receptors) and metabotropic receptors located on the postsynaptic neuron, thereby initiating excitatory signaling. This binding produces excitatory postsynaptic potentials and is crucial for synaptic plasticity and learning within reward circuits.
<b>GABAergic modulation</b> [410]	Gamma-Aminobutyric Acid (GABA)	GABA serves as the principal inhibitory neurotransmitter. At the synapse, GABA binds to GABAA receptors and GABAB receptors. This binding results in hyperpolarization of the postsynaptic cell, thereby providing inhibitory regulation that balances excitatory signals in the reward network.

## 5.1 The Human Reward Pathway

The brain’s reward system is broadly organized into two major anatomical pathways. The first is the medial forebrain bundle, which originates in the basal forebrain and projects through the midbrain, ultimately terminating in brainstem regions. The second is the dorsal diencephalic conduction system, which arises from the rostral portion of the medial forebrain bundle, traverses the habenula, and projects toward midbrain structures [407]. The feedback mechanisms and substances in the human brain are complex, involving a variety of neurotransmitters, hormones, and other molecules, which regulate brain function, emotions, cognition, and behavior through feedback mechanisms such as neurotransmitter systems and reward circuits. Feedback mechanisms can be positive (such as feedback in the reward system) or negative (such as inhibiting excessive neural activity). Well-known feedback substances [411] include dopamine, neuropeptides, endorphins, glutamate, etc.

Dopamine is a signaling molecule that plays an important role in the brain, affecting our emotions, motivation, movement, and other aspects [412]. This neurotransmitter is critical for reward-based learning, but this function can be disrupted in many psychiatric conditions, such as mood disorders and addiction. The mesolimbic pathway [406], a key dopaminergic system, originates from dopamine-producing neurons in the ventral tegmental area (VTA) and projects to multiple limbic and cortical regions, including the striatum, prefrontal cortex, amygdala, and hippocampus. This pathway plays a central role in reward processing, motivation, and reinforcement learning, and is widely recognized as a core component of the brain’s reward system. Neuropeptides are another important class of signaling molecules in the nervous system, involved in a variety of functions from mood regulation to metabolic control, and are slow-acting signaling molecules. Unlike neurotransmitters, which are limited to synapses, neuropeptide signals can affect a wider range of neural networks and provide broader physiological regulation. There is a significant cortical-subcortical gradient in the distribution of different neuropeptide receptors in the brain. In addition, neuropeptide signaling has been shown to significantly enhance the structure-function coupling of brain regions and exhibit a specialized gradient from

sensory-cognitive to reward-physical function [413]. Table 5 lists the common reward pathways in the human brain, the neurotransmitters they transmit, and the corresponding mechanisms of action, describing the basic framework of the human brain reward system.

## 5.2 From Human Rewards to Agent Rewards

Having examined the foundations of human reward pathways, we now turn to how artificial agents learn and optimize behavior through reward signals. While biological systems rely on complex neurochemical and psychological feedback loops, artificial agents operate using formalized reward functions designed to guide learning and decision-making. Though inspired by human cognition, agent reward mechanisms are structurally and functionally distinct. Understanding the analogies and disanalogies between these systems is crucial for aligning artificial behavior with human preferences.

In humans, rewards are deeply embedded in a rich web of emotional, social, and physiological contexts. They emerge through evolutionarily tuned mechanisms involving neurotransmitters like dopamine and are shaped by experiences, culture, and individual psychology. In contrast, artificial agents rely on mathematically defined reward functions that are externally specified and precisely quantified. These functions assign scalar or probabilistic feedback to actions or states, providing a signal for optimization algorithms such as reinforcement learning [3, 414].

One key distinction lies in the programmability and plasticity of agent rewards. Unlike human reward systems, which are constrained by biological architecture and evolutionary inertia, agent reward functions are fully customizable and can be rapidly redefined or adjusted based on task requirements. This flexibility enables targeted learning but also introduces design challenges—specifying a reward function that accurately captures nuanced human values is notoriously difficult.

Another important disanalogy concerns interpretability and generalization. Human rewards are often implicit and context-dependent, whereas agent rewards tend to be explicit and task-specific. Agents lack emotional intuition and instinctual drives; their learning depends entirely on the form and fidelity of the reward signal. While frameworks like reinforcement learning from human feedback (RLHF) attempt to bridge this gap by using preference data to shape agent behavior [12], such methods still struggle with capturing the full complexity of human goals, especially when preferences are intransitive, cyclical, or context-sensitive [321].

Moreover, attempts to borrow from human reward mechanisms—such as modeling intrinsic motivation or social approval—face limitations due to the absence of consciousness, embodiment, and subjective experience in artificial agents. Consequently, while human reward systems offer valuable inspiration, the design of agent reward functions must address fundamentally different constraints, including robustness to misspecification, adversarial manipulation, and misalignment with long-term human interests.

The following section will delve deeper into agent reward models, focusing on their design principles, evolution, and how these models selectively incorporate human-inspired insights to optimize artificial behavior within formal systems.

## 5.3 AI Reward Paradigms

Rewards also exist in intelligent agents, especially in reinforcement learning scenarios. Rewards are the core signal used to guide how intelligent agents act in the environment. They express feedback on the behavior of intelligent agents and are used to evaluate an action’s quality in a certain state, thereby affecting the decision-making of subsequent actions. Through continuous trial and error and adjustment, intelligent agents learn to choose behavioral strategies that can obtain high rewards in different states.

### 5.3.1 Definitions and Overview

In reinforcement learning, the reward model dictates how an agent is provided with feedback according to the actions it performs within its environment. This model plays a crucial role in guiding the agent’s behavior by quantifying the desirability of actions in a given state, thus influencing its decision-making.

**Formal Definition.** The agent’s interaction with its environment can be framed within the formalism of a Markov Decision Process (MDP) [415], which is represented as:

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma), \quad (5.1)$$

where:

- $\mathcal{S}$  denotes the state space, encompassing all possible states in the environment.
- $\mathcal{A}$  denotes the action space, which encompasses all actions available to the agent at any given state.
- $P(s'|s, a)$  defines the state transition probability. It represents the likelihood of transitioning to state  $s'$  after the agent takes action  $a$  in state  $s$ .
- $r(s, a)$  specifies the reward function, which assigns an immediate scalar reward received by the agent for executing action  $a$  in state  $s$ .
- $\gamma \in [0, 1]$  is the discount factor, which controls the agent's preference for immediate versus future rewards by weighting the contribution of future rewards to the overall return.

The reward function  $r(s, a)$  serves as a fundamental component in the formulation of the Agent Reward Model. It is mathematically represented as:

$$r(s, a) : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R} \quad (5.2)$$

This function returns a scalar reward based on the agent's current state  $s$  and the action  $a$  it selects. The scalar value  $r(s, a)$  is a feedback signal that indicates the immediate benefit (or cost) of the chosen action in the given state. This reward signal guides the agent's learning process, as it helps evaluate the quality of actions taken within specific contexts.

**Objective of the Agent Reward Model.** The agent's primary objective is to maximize its overall cumulative reward over time. This is typically achieved by selecting actions that yield higher long-term rewards, which are captured in the form of the return  $G_t$  at time step  $t$ , defined as the sum of future discounted rewards:

$$G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}, \quad (5.3)$$

where  $r_{t+k}$  denotes the reward received at time step  $t + k$ , and  $\gamma^k$  is the discount factor applied to rewards received at time step  $t + k$ . The agent aims to optimize its policy by maximizing the expected return over time.

At a higher level, the reward model can be classified into three categories based on the origin of the feedback signal: i) extrinsic reward, ii) intrinsic reward, iii) hybrid reward and iv) hierarchical model. Each of these categories can be further subdivided into smaller subclasses. Figure 5.2 illustrates different types of rewards. Next, we will explore these different types of reward in more detail, outlining the distinct features and applications of each type.

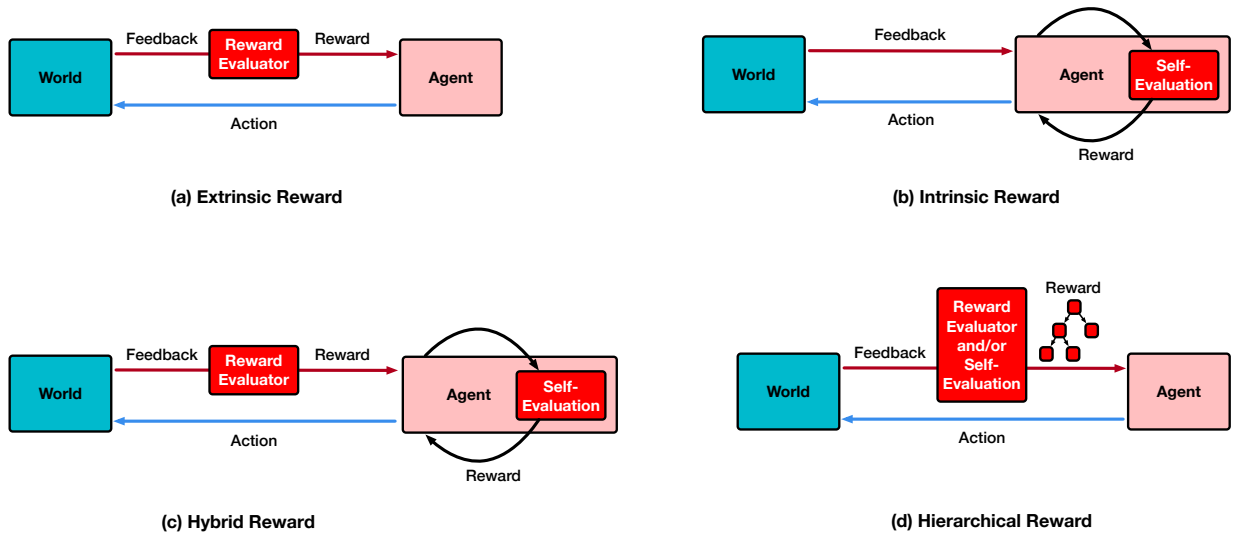


Figure 5.2: Illustration of different types of reward.

### 5.3.2 Extrinsic Rewards

Extrinsic rewards are externally defined signals that guide an agent’s behavior toward specific goals. In artificial learning systems, especially reinforcement learning, these signals serve as a proxy for success that shape the policy through measurable outcomes. However, the structure and delivery of these rewards significantly influence the learning dynamics, which present different trade-offs depending on how feedback is distributed.

**Dense Reward.** Dense reward signals provide high-frequency feedback, typically at every step or after each action. This frequent guidance accelerates learning by allowing agents to immediately associate actions with outcomes. However, dense feedback can sometimes incentivize short-sighted behavior or overfit to easily measurable proxies rather than deeper alignment.

For example, InstructGPT [43] uses human rankings of model outputs to provide continuous preference signals throughout fine-tuning, enabling efficient behavior shaping. Similarly, Cringe Loss [416] and its extensions [374] transform pairwise human preferences into dense training objectives, offering immediate signal at each comparison. Direct Reward Optimization (DRO) [367] further simplifies this paradigm by avoiding pairwise comparisons entirely, associating each response with a scalar score—making the reward signal more scalable and cost-effective. These methods exemplify how dense feedback facilitates fine-grained optimization but must be carefully designed to avoid superficial alignment.

**Sparse Reward.** Sparse rewards are infrequent and typically only triggered by major milestones or task completions. While they often reflect more meaningful or holistic success criteria, their delayed nature can make credit assignment more difficult, especially in complex environments.

PAFT [376] exemplifies this challenge by decoupling supervised learning and preference alignment, with feedback applied only at select decision points. This sparsity reflects a more global notion of success but increases the burden on optimization. Similarly, SimPO [377] uses log-probability-based implicit rewards without dense comparisons. The sparsity simplifies the training pipeline but can limit responsiveness to subtle preference shifts. Sparse reward systems thus tend to be more robust but demand stronger modeling assumptions or more strategic exploration.

**Delayed Reward.** Delayed rewards defer feedback until after a sequence of actions, requiring agents to reason about long-term consequences. This setup is essential for tasks where intermediate steps may be misleading or only make sense in retrospect. The challenge lies in attributing outcomes to earlier decisions, which complicates learning but encourages planning and abstraction.

Contrastive Preference Optimization (CPO) [384] trains models by comparing sets of translations rather than evaluating each one in isolation. The reward signal arises only after generating multiple candidates, reinforcing patterns across iterations. Nash Learning from Human Feedback [385] similarly delays feedback until the model identifies stable strategies through competitive comparisons. These methods leverage delayed rewards to push beyond surface-level optimization, aligning more with long-term goals at the cost of slower convergence and more complex training dynamics.

**Adaptive Reward.** Adaptive rewards evolve dynamically in response to the agent’s behavior or learning progress. By modulating the reward function such as increasing task difficulty or shifting reward targets, this approach supports continual improvement, especially in non-stationary or ambiguous environments. However, it introduces additional complexity in reward design and evaluation.

Self-Play Preference Optimization (SPO) [386] adapts rewards based on self-play outcomes, using social choice theory to aggregate preferences and guide learning. This approach allows the system to refine itself by evolving internal standards. f-DPO [373] builds on this idea by introducing divergence constraints that adapt the reward landscape during training. By tuning alignment-diversity trade-offs dynamically, these methods enable robust preference modeling under uncertainty, though they require careful calibration to avoid instability or unintended bias.

### 5.3.3 Intrinsic Rewards

Intrinsic rewards serve as internally generated signals that motivate agents to explore, learn, and improve, independent of external task-specific outcomes. These rewards are often structured to promote generalization, adaptability, and self-directed skill acquisition—qualities critical for long-term performance in complex or sparse-reward environments. Different intrinsic reward paradigms focus on fostering distinct behavioral tendencies within agents.

**Curiosity-Driven Reward.** This reward encourages agents to reduce uncertainty by seeking novel or surprising experiences. The key concept is to incentivize the agent to explore novel states where prediction errors are significant. This paradigm excels in sparse-reward settings by promoting information acquisition when external guidance is limited. For example, Pathak et al. [387] leverage an inverse dynamics model to predict the outcome of actions, creating a feedback loop that rewards novelty. Plan2Explore [389] extends this further by incorporating forward planning to

actively target areas of high epistemic uncertainty, thereby enabling faster adaptation to unseen environments. While effective at discovery, curiosity-driven methods can be sensitive to noise or deceptive novelty without safeguards.

**Diversity Reward.** Diversity reward shifts focus from novelty to behavioral heterogeneity, encouraging agents to explore a wide range of strategies rather than converging prematurely on suboptimal solutions. This approach is particularly useful in multi-agent or multimodal settings, where strategic variety enhances robustness and collective performance. LIIR [390] exemplifies this by assigning personalized intrinsic signals to different agents, driving them toward distinct roles while maintaining shared objectives. Diversity-driven exploration fosters broader policy coverage but may require careful balancing to avoid destabilizing coordination or goal pursuit.

**Competence-Based Reward.** Competence-based reward aims to foster learning progress by rewarding improvements in the agent’s task proficiency. This reward adapts dynamically as the agent grows more capable, which creates a self-curriculum that supports continual skill acquisition. Skew-Fit [392] facilitates this through entropy-based goal sampling, encouraging agents to reach diverse states while maintaining challenge. CURIOUS [391] further automates curriculum generation by selecting goals that maximize learning progress over time. Competence-based methods are well-suited for open-ended environments, though they often require sophisticated estimation of progress and goal difficulty.

**Exploration Reward.** Exploration reward directly incentivizes the agent to engage with under-explored states or actions, which emphasize breadth over depth in environment interaction. Unlike curiosity, which focuses on unpredictability, exploration reward often targets coverage or novelty relative to the agent’s visitation history. RND [394] exemplifies this by rewarding the prediction error of a randomly initialized network, pushing the agent toward unfamiliar states. This approach helps prevent premature convergence and encourages robustness, though it may lack focus if not paired with meaningful learning objectives.

**Information Gain Reward.** Information gain reward formalizes exploration as a process of uncertainty reduction, which guides agents to take actions that yield the highest expected learning. This reward is grounded in information theory and is especially powerful in model-based or reasoning-intensive tasks. CoT-Info [397] applies this to language models by quantifying knowledge gain at each reasoning step, optimizing sub-task decomposition. VIME [398] similarly employs Bayesian inference to reward belief updates about environmental dynamics. By explicitly targeting informational value, these methods offer principled exploration strategies, though they often incur high computational cost and require accurate uncertainty modeling.

### 5.3.4 Hybrid Rewards

Hybrid reward frameworks integrate multiple sources of feedback, most commonly intrinsic and extrinsic rewards, to enable more balanced and adaptive learning. By combining the exploratory drive of intrinsic rewards with the goal-directed structure of extrinsic rewards, these systems aim to improve both sample efficiency and generalization. This paradigm is especially beneficial in complex environments or open-ended tasks, where pure reliance on either feedback type may be insufficient.

A core advantage of hybrid rewards is their capacity to resolve the exploration-exploitation trade-off dynamically. For instance, Xiong et al. [403] combine intrinsic exploration with extrinsic human feedback within the context of RLHF. Using a reverse-KL regularized contextual bandit framework, they facilitate strategic exploration while aligning the agent’s actions with human preferences. The method integrates intrinsic and extrinsic rewards through an iterative DPO algorithm and multi-step rejection sampling, optimizing exploration and alignment without compromising efficiency.

### 5.3.5 Hierarchical Rewards

Hierarchical reward architectures decompose complex objectives into layered subgoals, each associated with distinct reward signals. This structure mirrors the hierarchical organization of many real-world tasks, allowing agents to coordinate short-term decisions with long-term planning. By assigning lower-level rewards to immediate actions and higher-level rewards to abstract goals, agents can learn compositional behaviors that scale more effectively to complex environments.

In language modeling, Token-level Direct Preference Optimization (TDPO) [405] illustrates this principle by aligning LLMs through fine-grained token-level rewards derived from preference modeling. Using forward KL divergence and the Bradley-Terry model, TDPO simultaneously refines local choices and global coherence, improving alignment with nuanced human preferences. The hierarchical reward process here is not merely a structural design but a functional one: reinforcing both micro-decisions and macro-outcomes in a coordinated fashion.



More generally, hierarchical rewards can serve as scaffolding for curriculum learning, where agents progressively learn from simpler subtasks before tackling the overarching objective. In LLM agents, this might mean structuring rewards for subcomponents like tool-use, reasoning chains, or interaction flows, each of which contributes to broader task success.

## 5.4 Summary and Discussion

### 5.4.1 Interaction with Other Modules

In intelligent systems, reward signals function not only as outcome-driven feedback but as central regulators that interface with core cognitive modules such as perception, emotion, and memory. In the context of LLM-based agents, these interactions become particularly salient, as modules like attention, generation style, and retrieval memory can be directly influenced through reward shaping, preference modeling, or fine-tuning objectives.

**Perception.** In LLM agents, perception is often realized through attention mechanisms that prioritize certain tokens, inputs, or modalities. Reward signals can modulate these attention weights implicitly during training, reinforcing patterns that correlate with positive outcomes. For example, during reinforcement fine-tuning, reward models may upweight specific linguistic features—such as informativeness, factuality, or politeness—causing the model to attend more to tokens that align with these traits. This parallels how biological perception prioritizes salient stimuli via reward-linked attentional modulation [417]. Over time, the agent internalizes a perception policy: not merely “what is said,” but “what is worth paying attention to” in task-specific contexts.

**Emotion.** Though LLMs do not possess emotions in the biological sense, reward signals can guide the emergence of emotion-like expressions and regulate dialogue style. In human alignment settings, models are often rewarded for generating responses that are empathetic, polite, or cooperative—leading to stylistic patterns that simulate emotional sensitivity. Positive feedback may reinforce a friendly or supportive tone, while negative feedback suppresses dismissive or incoherent behavior. This process mirrors affect-driven behavior regulation in humans [418], and allows agents to adapt their interaction style based on user expectations, affective context, or application domain. In multi-turn settings, reward-modulated style persistence can give rise to coherent personas or conversational moods.

**Memory.** Memory in LLM agents spans short-term context (e.g., chat history) and long-term memory modules such as retrieval-augmented generation (RAG) or episodic memory buffers. Reward signals shape how knowledge is encoded, reused, or discarded. For instance, fine-tuning on preference-labeled data can reinforce certain reasoning paths or factual patterns, effectively consolidating them into the model’s internal knowledge representation. Moreover, mechanisms like experience replay or self-reflection—where agents evaluate past outputs with learned reward estimators—enable selective memory reinforcement, akin to dopamine-driven memory consolidation in biological systems [419]. This allows LLM agents to generalize from prior successful strategies and avoid repeating costly errors.

In general, reward in LLM-based agents is not a passive scalar signal but an active agent of behavioral shaping. It modulates attention to promote salient features, guides stylistic and affective expression to align with human preferences, and structures memory to prioritize useful knowledge. As agents evolve toward greater autonomy and interactivity, understanding these cross-module reward interactions will be essential for building systems that are not only intelligent, but also interpretable, controllable, and aligned with human values.

### 5.4.2 Challenges and Directions

Although extensive research has been conducted on various reward mechanisms, several persistent challenges remain. One fundamental issue is reward sparsity and delay. In many real-world scenarios, reward signals are often infrequent and delayed, making it difficult for an agent to accurately attribute credit to specific actions. This, in turn, increases the complexity of exploration and slows down the learning process.

Another significant challenge is the potential for reward hacking. Agents, in their pursuit of maximizing rewards, sometimes exploit unintended loopholes in the reward function. This can lead to behaviors that diverge from the intended design goals, particularly in complex environments where optimization objectives may not always align with the true task requirements.

Moreover, the process of reward shaping presents a delicate balance. While shaping rewards can accelerate learning by guiding an agent toward desired behaviors, excessive or poorly designed shaping may lead to local optima, trapping the agent in suboptimal behaviors. In some cases, it may even alter the fundamental structure of the original task, making it difficult for the agent to generalize to other scenarios.

Many real-world problems are inherently multi-objective in nature, requiring agents to balance competing goals. Under a single reward function framework, finding the right trade-offs between these objectives remains an open problem. Ideally, a hierarchical reward mechanism could be designed to guide learning in a structured, step-by-step manner. However, constructing such mechanisms effectively is still a challenge.

Finally, reward misspecification introduces further uncertainty and limits generalization. Often, a reward function does not fully capture the true task goal, leading to misalignment between the agent's learning objective and real-world success. Additionally, many reward functions are tailored to specific environments and fail to generalize when conditions change or tasks shift, highlighting the need for more robust reward models.

Addressing these challenges requires novel approaches. One promising direction is to derive implicit rewards from standard examples or outcome-based evaluations, which can help mitigate reward sparsity issues. Additionally, decomposing complex tasks into hierarchical structures and designing rewards from the bottom up can offer a more systematic approach, even in multi-objective settings. Furthermore, leveraging techniques such as meta-learning and meta-reinforcement learning can enhance the adaptability of reward models, allowing agents to transfer knowledge across tasks and perform effectively in diverse environments. By exploring these avenues, we can move toward more reliable and scalable reward mechanisms that better align with real-world objectives.